Exercise 2

A CPU has frequency $f = 1$ GHz. It can do one load, one multiplication and one addition per clock cycle.

The memory bus has bandwidth $B = 3.2$ GBytes / s.

The latency to load one cache line from memory is
$T_\ell = 100 \text{ clock cycles} = \dfrac{100}{1 \text{ GHz}} = 100 \cdot 10^{-9} \text{ s} = 100 \text{ ns}$.

One cache line can hold four double precision objects, i.e the length of a cache line is $L_c = 4 \cdot 8 \text{ Bytes} = 32 \text{ Bytes}$.

```
double s = 0;
for (int i = 0; i < N; i++){
  s = s + A[i]*B[i];
```

a)

The code inside the loop consists of three loads (A[i], B[i] and s), one multiplication, one addition and one store (s).

As the data item s is declared and used repeatedly in each loop, I'm assuming that it always resides in a register and does not have to be brought from memory each time it is used. Thus I'm assuming the first cache line contains A[0], B[0], A[1[, B[1], the second cache line contains A[2], B[2], A[3], B[3] and so on.

First loop

The time to bring the first cache line from memory is $T = T_\ell + \dfrac{L_c}{B}$. At this point A[0], B[0], A[1[, B[1] all resides in registers.

Then the CPU will use one clock cycle to load A[i], then a single clock cycle to load B[i], multiply A[i]*B[i], add s + A[i]*B[i] and store s. I.e executing the code takes two clock cycles.

Second loop

As A[1] and B[1] already resides in registers, they do not have to be brought from memory. Thus the CPU can immediately execute the code inside the loop which takes two clock cycles.

Third loop

Repeat of first loop.

The time to execute the first and second loop is $T + 4 \cdot \frac{1}{f} = T_\ell + \frac{L_c}{B} + \frac{4}{f}$. As a single loop contains two floating-point operations, the expected performance is

$$\text{FLOPS} = \frac{4 \text{ flops}}{T_\ell + \frac{L_c}{B} + \frac{4}{f}} = \frac{4 \text{ flops}}{100 \text{ ns} + \frac{32 \text{ Bytes}}{3.2 \text{ GBytes/s}} + \frac{4}{1 \text{ GHz}}} = \frac{4 \text{ flops}}{100 \text{ ns} + 10 \text{ ns} + 4 \text{ ns}} = \frac{4 \text{ flops}}{114 \text{ ns}}$$

$$= \frac{4}{114} \text{ Gflops} / \text{s} \approx 35.1 \text{ Mflops} / \text{s}$$

b)

$$P = 1 + \frac{T_\ell}{L_c / B} = 1 + \frac{100 \text{ ns}}{10 \text{ ns}} = 1 + 10 = 11$$

c)

Twice as long:

$$P = 1 + \frac{T_\ell}{2L_c / B} = 1 + \frac{100}{20} = 1 + 5 = 6$$

Four times as long:

$$P = 1 + \frac{100}{40} = 3.5$$

d)

The performance is now

$$\text{FLOPS} = \frac{4\text{ flops}}{\frac{L_c}{B} + \frac{4}{f}} = \frac{4\text{ flops}}{10\text{ ns} + 4\text{ ns}} = \frac{4}{14}\text{ Gflops}\,/\,\text{s} \approx 285.7\text{ Mflops}\,/\,\text{s}$$